DOI：10.11830/ISSN.1000-5013.202103034

注意力叠加与时序特征融合的目标检测方法

吴雨泽，聂卓赟，周长新

（华侨大学 信息科学与工程学院，福建 厦门 361021）

摘要： 提出一种基于注意力叠加与时序特征融合的目标检测方法.在端到端目标检测(DETR)网络的基础上，依据注意力机制特性，使用注意力权重叠加的方式提取目标物像素级标识，用于实例轨迹的划分.为使目标检测与轨迹跟踪协同作用，通过时序特征融合的方式融合之前轨迹跟踪信息，调整当前帧目标检测效果，从而充分利用视频载体提供的时间维度信息.在公开数据集上，对文中方法进行验证，结果表明：文中方法能有效识别被遮挡的目标物，具有较强鲁棒性.

关键词： 目标检测网络；注意力机制；轨迹跟踪；时序特征

中图分类号： TP 183；TP 391.4 **文献标志码：** A **文章编号：** 1000-5013(2022)05-0650-08

Object Detection Method of Attention Superposition and Temporal Feature Fusion

WU Yuze, NIE Zhuoyun, ZHOU Changxin

(College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China)

Abstract: An object detection method of attention superposition and temporal feature fusion is proposed. Based on the end-to-end object detection (DETR) network, attention weight superposition is used to extract the object pixel-level identification for the division of the instance trajectory according to the characteristics of the attention mechanism. In order to cooperate the object detection and trajectory tracking, the previous track tracking information is fused by the temporal feature fusion to adjust the effect of current frame object detection, so as to make full use of the temporal dimension information provided by the video carrier. The proposed method is experimentally tested on the public data set. The results show that the method in this paper can effectively detect the blocked object and has stronger robustness.

Keywords: object detection network; attention mechanism; trajectory tracking; temporal feature

在先进驾驶辅助系统与自动驾驶技术中，图像视频处理是重要的一环，通过机器视觉对行人、车辆等交通环境进行目标检测[1]，以确保驾驶安全.然而，实际行车环境的复杂性(光照、雨雪天气、道路杂物、道路拥挤等)常导致检测目标被遮挡，从而引起安全隐患，特别是在高速行驶下的目标丢失极具危险性.因此，研究适用于短暂目标遮挡的目标检测算法具有重要意义.

近年来，随着深度学习的发展，目标检测算法取得了很大的突破，如两阶段的 FasterR-CNN[2-4] 系

收稿日期： 2021-03-21

通信作者： 聂卓赟(1983-)，男，副教授，博士，主要从事鲁棒控制及非线性系统的研究.E-mail：yezhuyun2004@sina.com.

基金项目： 国家自然科学基金资助项目(61403149)；福建省自然科学基金资助项目(2019J01053)

列、一阶段的 SSD 系列[5-9]和 YOLO 系列[10-13]. 特别是 2020 年 Facebook AI 提出直接将 transformer 架构适配到视觉领域的端到端目标检测(DETR)网络[14]，将注意力机制在图像中的应用带到了台前，并取得了前所未有的检测精度. 国内商汤科技的也进一步肯定了注意力机制的作用[15-17].

目前，大多数应用场景是以视频作为信息的载体[19]，除了常规视觉信息外，视频还提供了额外的时间维度信息. 针对视频进行目标检测，通常可以进行两类识别计算：一类是对每一帧进行目标检测；另一类是利用时间维度信息，将被检测的目标物跨帧链接成轨迹[18-25]，并利用生成的轨迹进行下次目标状态预测和跟踪. 在检测任务中，为充分利用时间维度信息，本文将检测与跟踪集成于一体，以 DETR 网络为主干，通过时间序列信息增强目标检测的效果.

1　融合上下文信息的增强型目标检测网络

为融合视频时间维度信息以增强当前帧的检测效果，在单帧网络的基础上，提出依据网络注意力特性的增强型目标检测网络. 为形式化描述网络运行过程，给定一段视频序列 $F_t, t=0,\cdots,T$，目的是检测所有出现的目标物. 设 $0\sim t$ 时刻所有的跟踪目标实例为 $\boldsymbol{D}_t=\{\boldsymbol{d}_j^t,\boldsymbol{c}_j^t\}$，其中，$\boldsymbol{d}_j^t$ 表示某一跟踪实例 j，$j=1,\cdots,m$；$\boldsymbol{c}_j^t$ 表示实例 j 所属类别. 对于视频序列的每一帧 $I_{t_k}, t_k\leqslant t$，目标检测输出为 $b_{k^t}^{t_k}\in B_t, t_k\leqslant t$，表示第 t_k 帧的第 k 个包围框，且 $\boldsymbol{d}^t=\{b_{k^t}^{t_k}\}, t_k\leqslant t$，跟踪实例由多帧检测目标对应匹配构成. 若第 t 帧的检测结果网络关注的图像上空间位置信息(注意力)用 σ_t^b 表示，该空间位置下提取的作为标识“本帧检测”目标的特征则为 $\sigma_t^b(F_t)$，相对的“标识跟踪”实例的特征用 $\sigma_t^d(F_t)$ 表示. 因此，网络整体可用形式化语言表述.

算法 1：融合跟踪目标信息的目标检测算法

输入：video frames F_t

输出：bounding box $B_t, t=0,\cdots,T$

$B_0=\mathrm{Detection}_{\mathrm{DETR}}(F_0)$

$D_0=B_0$

For t=1 to T

　$\omega=\mathrm{CosineSimilarity}(\sigma_t^b(F_t),\sigma_{t-1}^d(F_t))$

　$B_t=\mathrm{EnhancedDetect}(\omega,F_t,D_{t-1})$

　$B_t=\mathrm{NMS}(B_t)$

$D_t=\mathrm{Binary\ Graph\ Matching}(B_t,D_{t-1})$

1.1　整体框架

整体框架，如图 1 所示. 图 1 中：CNN 为卷积神经网络；FFN 为前馈神经网络；NMS 为非极大值抑制层；re-id 为重检测；$\boldsymbol{x}'$ 为提取的特征序列；$\boldsymbol{x}$ 为叠加位置编码后的图像底层特征信息；$\boldsymbol{y}$ 为对应位置的特征向量；$\boldsymbol{z}$ 为输出的特征序列；α,θ,γ 为可学习参数；$\boldsymbol{\omega}$ 为相似度.

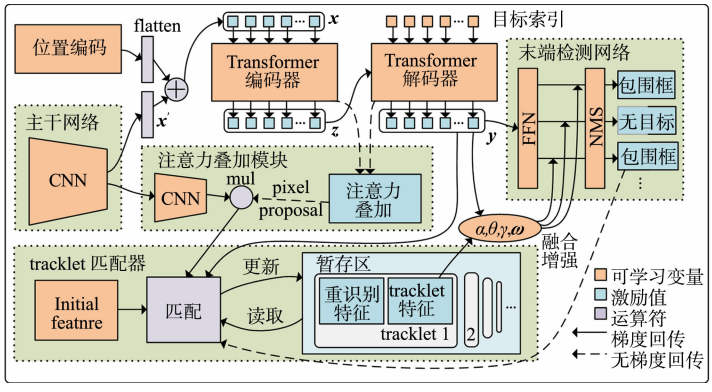

图 1　整体框架

Fig. 1　Overall framework

DETR 采用 resnet-50[15]作为主干网络，该网络以 Block 作为基本单位，每个 Block 包含两层 1×1 的卷积层和一层 3×3 的卷积层，外加跳接层传输残差，其中，1×1 卷积层主要用于减少通道数，降低网

络整体参数量. 主干网络一共包含48个Block用于提取图像底层特征,而后展平为序列 $\boldsymbol{x}'$,叠加位置编码后的 $\boldsymbol{x}$ 输入Transformer[16]网络. Transformer网络分为Encoder与Decoder两大模块. 不同于以往CNN获取图像信息的方式,Encoder采用文献[16]所提出的注意力机制,主干网络提取的特征序列经过自注意力层与全连接层(共有6组Encoder串接),可以提取图像中长距离相关信息作为输出的特征序列 $\boldsymbol{z}$,克服了CNN过于关注局部信息的问题. Decoder的输入为固定数目的object query序列(每个序列元素代表图像某个位置的编码,可学习)及Encoder输出的从图像中所提取到的特征序列 $\boldsymbol{z}$,输出为图像上该object query对应位置的特征向量 $\boldsymbol{y}$. 经过两层FFN和NMS[17]后,得到图像上对应位置包围框的定位数值序列 $\boldsymbol{b}$ 及包围框内存在目标物的概率值 $P(c|\boldsymbol{b})$,其中,c 表示类别.

为解决视频信息中常存在的模糊、遮挡问题,提出基于注意力叠加的标识提取模块,将叠加的Transformer网络注意力 σ^b 作为每个输出目标的空间信息,用于抽取图像上对应空间的底层特征. 通过计算跟踪目标(上、下帧所有检测目标经二分图匹配所得)的相似度 $\boldsymbol{\omega}$,将已有跟踪目标的特征信息融入本帧检测,增强本帧图像目标检测效果.

1.2 注意力机制

Attention机制最早在视觉领域提出,Google Mind采用循环神经网络(RNN)模型结合Attention机制完成图像分类需求[17]. 而后,Bahdanau等[25]又将其引入到神经语言程序学(NLP)领域,采用Seq2Seq融合Attention机制进行机器翻译. 将Attention机制推向研究热点的是Google机器翻译团队,Vaswani[16]提出Transformer网络结构,完全抛弃RNN,CNN等传统结构,仅仅依靠Attention机制进行翻译任务,并取得惊人的效果.

Transformer网络结构,如图2所示. 图2中:a为解码器的展开;b为Transformer网络简图;c为中的编码器模块,通过上、下文(即寻找源句中与之相关的词语,称为自注意力,即每个词汇对其他词汇

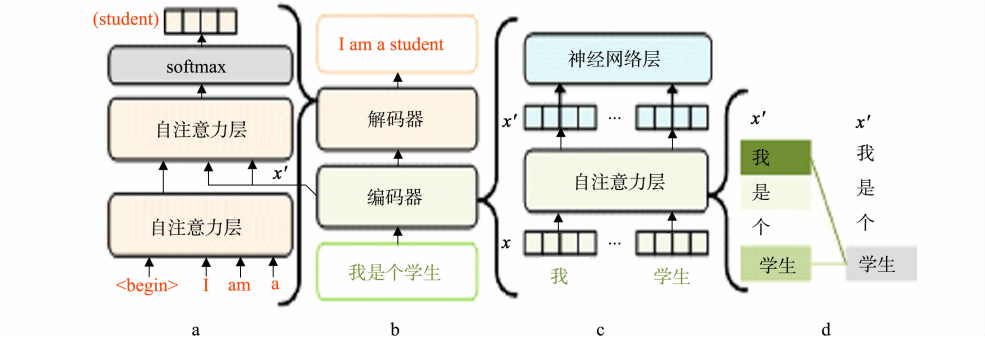

图2 Transformer网络结构

Fig. 2 Network structure of Transformer

关注度的权重)将每个词汇编码为中间向量 $\boldsymbol{r}$;d为自注意力层的展开. 求某词汇相对另外一个词汇的"关注度"时,注意力权重 $\boldsymbol{\omega}$ 最终输出 $\boldsymbol{r}=\boldsymbol{\omega}\times\boldsymbol{v}$. 以此逐个计算每个词汇相对源句中其他词汇的注意力,使 $\boldsymbol{r}$ 中每个语义编码都包含某个词汇在特定上、下文表示的语义信息,即

$$\boldsymbol{\omega}_{\mathrm{Att}}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})=\mathrm{softmax}(\boldsymbol{q}\times\boldsymbol{k}). \tag{1}$$

式(1)中:$\boldsymbol{q}$ 为该词汇的索引信息;$\boldsymbol{k}$ 为匹配信息,$\boldsymbol{v}$ 为词汇语义信息.

通过引入注意力机制,克服了RNN,CNN在计算序列信息上存在的窗口问题. 序列上每个元素之间的距离不再成为影响结果的重要因素,使长期记忆变为可能. 解码器同样使用注意力机制,完成从语义信息编码到目标语言编码的变换,区别只在于解码器注意力层使用编码器输出 $\boldsymbol{k}$,$\boldsymbol{v}$,不再赘述.

1.3 基于注意力叠加的像素级标识的提取模块

2020年,Facebook AI提出直接将transformer架构适配到视觉领域的DETR网络,不再采用传统的基于预先生成的锚定框回归候选框误差的策略,直接从图像特征并行地回归候选框与类别,最大特点是使用注意力机制Transformer网络中Encoder层自注意力权重热点图,如图3所示. 图3中:(8,13),(9,25),(9,3),(8,32)为编码器在特征图上的4个位置.

由图3可知:在DETR网络中,Encoder每层的注意力有集中于目标实例的特性,而在Decoder各

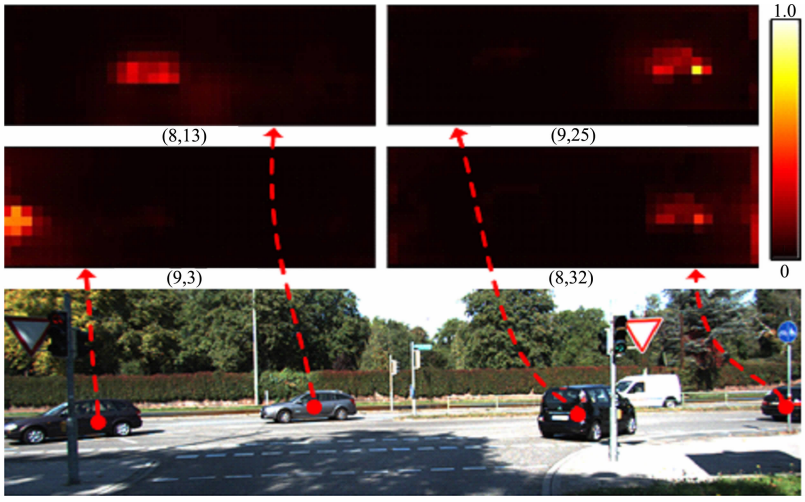

图 3　Transformer 网络中 Encoder 层自注意力权重热点图

Fig. 3　Self-attention weight hot spot diagram of encoder layer in transformer network

层的注意力则分散于每个目标实例的边缘处. 因此，依据该特性提出基于注意力叠加的标识提取模块，通过叠加 Encoder 与 Decoder 每个检测实例的注意力权重，获取在预测该目标时网络重点关注的空间位置信息，用于提取像素级别的网络低维度特征. Transformer 网络中 Encoder 层各通道自注意力权重的叠加，如图 4 所示. 图 4 中：a 为解码器注意力叠加；b 为编码器自注意力叠加.

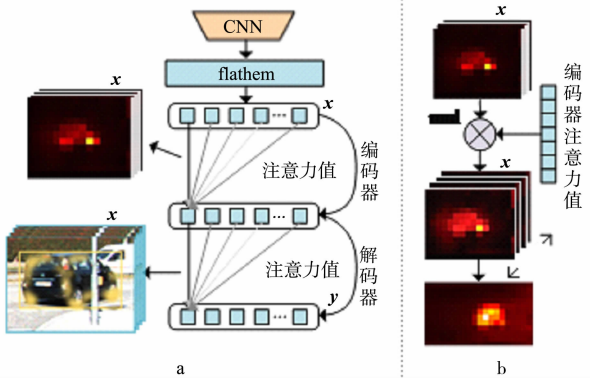

图 4　Transformer 网络中 Encoder 层各通道自注意力权重的叠加

Fig. 4　Superposition of Self-attention weight of each channel in the encoder layer in transformer network

由图 4 中的 a 可知：原图像经过主干网络输出为 $w\times h$ 的 featuremap，将其展平成长度为 $n=w\times h$ 的一维向量 $\boldsymbol{x}$，则 Encoder 层注意力权重矩阵为 $\boldsymbol{N}^{n\times n}$，每一行代表某像素点对与其他像素点的注意力值，共 n 行，即有 n 幅注意力权重热点图.

同理，Decoder 层注意力权重矩阵为 $\boldsymbol{M}^{m\times n}$，$m$ 为 object query 序列长度（有 100 个预测结果），每一行代表某个输出对 Encoder 输出 $\boldsymbol{r}$ 的注意力权重. 设 $\boldsymbol{\sigma}^{m\times n}$ 代表 m 个预测结果所关注的 featuremap 上的空间位置信息，则图 4 中 b 的第 i 个预测结果的权重信息为

$$\boldsymbol{\sigma}^{i}=\sum_{j=0}^{n-1}\boldsymbol{N}_{j}\cdot M_{i,j}. \tag{2}$$

对于视频序列 $\boldsymbol{F}_t$ 的第 t 帧，预测结果包围框为 K 的目标物，网络提取的目标物标识为

$$\boldsymbol{\sigma}^{K}(\boldsymbol{F}_t)=\boldsymbol{\sigma}_{K}\cdot\boldsymbol{x}_t=\sum_{j=0}^{n-1}\boldsymbol{N}_{j}\cdot M_{K,j}\cdot\boldsymbol{x}_t\,. \tag{3}$$

式(3)中：$\boldsymbol{x}_t$ 为主干网输出的特征图.

1.4　时序特征的融合模块

网络通过之前帧的目标检测与跟踪所提供的信息增强本帧目标检测效果，也即在构建的网络在进行目标检测时，不仅使用了本帧里的图像信息，还使用了从前些帧中恢复的 tracklet 信息，即增强型目

标检测网络融合了上、下文信息.

对于给定的一段视频 $\boldsymbol{F}_t$,$\boldsymbol{b}_i^t$ 表示第 t 帧的第 i 个候选框,$\boldsymbol{d}_j^{t-1}$ 表示从 0～(t－1)时刻,由各个时刻同一目标的检测框所组成的跟踪目标实例 j,$\boldsymbol{D}_t$ 表示所有的跟踪目标,$P(c|\boldsymbol{b}_i^t,\boldsymbol{D}_t)$表示在已有的跟踪目标 $\boldsymbol{D}_t$ 前提下,第 t 帧第 i 个预测框的类型为 c 的概率.一般而言,候选框与已有的某跟踪目标越像,就越有可能采用同样的类别标签.基于此,网络在预测候选框标签的时候,考虑各个已有跟踪目标的相关性,有

$$P(c|\boldsymbol{b}_i^t,\boldsymbol{D}_t)=\sum_{j=0}^{m}\boldsymbol{\omega}(\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1})P(c\mid\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1}). \tag{4}$$

式(4)中:$\boldsymbol{\omega}$ 是本帧第 i 个候选框 $\boldsymbol{b}_i^t$ 与第 j 个跟踪目标 $\boldsymbol{d}_j^{t-1}$ 的相似度,计算公式为

$$\left.\begin{aligned}&\boldsymbol{\omega}'(\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1})=\text{Min Max Normal}(\theta^{\gamma\cdot\cos\text{Similarity}(\sigma_K(\boldsymbol{I}_t),\sigma_d(\boldsymbol{I}_{t-1}))}),\\&\boldsymbol{\omega}(\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1})=\begin{cases}\boldsymbol{\omega}'(\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1}), & \boldsymbol{\omega}'(\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1})>0.6,\\0, & \text{其他}.\end{cases}\end{aligned}\right\} \tag{5}$$

式(5)中:参数 θ,γ 初始值分别为 $\theta=20.0$,$\gamma=8.0$.

$P(c|\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1})$由候选框 i 与实例 j 所提供的信息决定,首先,计算本帧候选框标识与跟踪目标标识的余弦相似度.其次,通过指数函数增加数据的区分度,最后,归一化到 0～1 范围内,并截断于 0.6.其表达式为

$$P(c|\boldsymbol{b}_i^t,\boldsymbol{d}_j^{t-1})=P(c|\boldsymbol{b}_i^t)+\alpha P(c|\boldsymbol{d}_j^{t-1}). \tag{6}$$

式(6)中:α 的初始值为 0.5.

跟踪实例的类别由每帧中同一实例标签所决定,其中,$\boldsymbol{b}_k^t$ 是匹配给跟踪目标 $\boldsymbol{d}_j^t$ 的包围框,即

$$P(c|\boldsymbol{d}_j^t)=\frac{P(c|\boldsymbol{b}_k^t,\boldsymbol{D}_{t-1})+\beta P_{\text{tr}}(c|\boldsymbol{d}_j^{t-1})\text{len}(\boldsymbol{d}_j^{t-1})}{1+\beta\text{len}(\boldsymbol{d}_j^{t-1})}. \tag{7}$$

包围框与跟踪目标的匹配有多种方法[10],采用基本的 KM 算法进行匹配,代价矩阵 $\boldsymbol{C}$ 为

$$\left.\begin{aligned}&\boldsymbol{C}=\boldsymbol{C}_{\text{box}}+\boldsymbol{C}_{\text{cla}}+\boldsymbol{C}_{\text{gio}},\\&\boldsymbol{C}_{i,j}^{\text{box}}=\|\boldsymbol{b}_i^{\text{pre}}-\boldsymbol{b}_j^{\text{tra}}\|_1,\\&\boldsymbol{C}_{i,j}^{\text{cla}}=1-P(c_j|\boldsymbol{b}_i^{\text{pre}}),\\&\boldsymbol{C}_{i,j}^{\text{gio}}=\text{Max Min Normal}(1-\text{IOU}(\boldsymbol{b}_i^{\text{pre}},\boldsymbol{b}_j^{\text{tra}})).\end{aligned}\right\} \tag{8}$$

本帧所预测的目标为以最小代价匹配已有跟踪实例,设定最大代价值为 2.0,超过此值忽略匹配.本帧预测目标匹配成功的实例根据式(7)更新已有跟踪实例数据,已有跟踪实例在本帧未被匹配者,则从缓存中删除.即

$$\boldsymbol{C}_{i,j}^{\text{gio}}=\text{Max Min Normal}(1-\text{IOU}(\boldsymbol{b}_i^{\text{pre}},\boldsymbol{b}_j^{\text{tra}})).$$

2 实验部分

实验平台硬件配置如下:Intel(R) Xeon(R) CPU E5-2623 v4@2.60 GHz;内存 32 GB;TITAN Xp 型显卡,12 GB.软件配置如下:Ubuntu18.04LTS,CUDA11.2,python3.6,pytorch1.6.0.为验证文中方法,采用 KITTI 数据集对不同方法进行对比.KITTI 数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办,是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集.该数据集用于评测立体图像、光流、视觉测距、3D 物体检测和 3D 跟踪等.KITTI 数据集包含市区、乡村和高速公路等场景采集的真实图像数据,每张图像中最多达 15 辆车和 30 个行人,还有各种程度的遮挡与截断.同时,KITTI 数据集包含多个类别的标签,但大部分类别数据量较少,文中取数据最多的 Car,Pedestrian 与 Cyclist 三项.

2.1 模型训练

为减短模型训练时间,部分网络初始权重采用在 ImageNet 上预训练好的 DETR 模型参数,并冻结主干网络权重,不参与学习.损失函数 loss 为

$$\left.\begin{aligned}&\text{loss}=\rho\cdot\text{loss}_{\text{lab}}+\zeta\cdot\text{loss}_{\text{box}},\\&\text{loss}_{\text{lab}}=\text{Cross Entropy}(P(c|B^{\text{pre}}),P(c|B^{\text{tru}})),\\&\text{loss}_{\text{box}}=\sum_{i=1}^{n}\frac{|\boldsymbol{b}_i^{\text{pre}}-\boldsymbol{b}_i^{\text{tru}}|}{n}.\end{aligned}\right\} \tag{9}$$

式(9)中：ρ，ζ 为超参数，均取 1；B^{pre} 为预测的包围框集合，B^{tru} 为包围框标签集合.

训练采用 Adam 优化算法，初始学习率为 0.000 1，权重衰减参数为 1×10^{-5}，学习率衰减策略为 StepLR，StepSize 为 200. 输入图像分辨率 1 242 px×375 px，为体现网络在遮挡条件下融合上、下文信息预测的能力，除通常的归一化等图像预处理外，还对目标进行随机的遮挡. 训练集图片总数为 7 121，随机抽取 3 张连续帧作为一个训练单元，则每个 epoch 有 7 118 个训练单元. 网络训练损失函数曲线，如图 5 所示. 图 5 中：loss 为损失函数. 由图 5 可知：网络在快速下降后，逐渐趋于平稳.

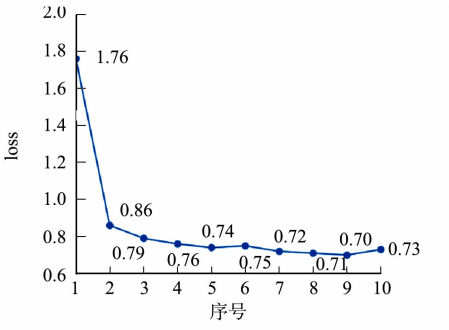

图 5　网络训练损失函数曲线

Fig. 5　Loss function curve of network training

2.2　检测结果

被遮挡目标的检测效果图，如图 6 所示. 图 6 中：每个样本包含连续的 3 帧图像，并在第 3 帧人为添加遮挡物；红色框为目标检测效果；蓝色框为原 DETR 网络目标检测效果. 为排除训练次数带来的干扰，测试的两模型中相同网络部分具有同样的权重.

(a) 样本 1　　(b) 样本 2

(c) 样本 3　　(d) 样本 4

图 6　被遮挡目标的检测效果图

Fig. 6　Detection effect diagram of occluded target

由图 6 可知：目标物无遮挡的情况下，两网络表现效果相近；而对于被遮挡的目标物(遮挡物为实心黑色圆，遮挡比(η)为圆半径与包围框宽度比)，文中网络具有更好的效果.

改进前、后目标检测精度数据，如表 1 所示. 表 1 中：AP 为精度.

表1 改进前、后目标检测精度数据

Tab.1 Target detection accuracy dates of improving before and after

η	检测方法	AP(Car)/%	AP(Pedestrian)/%	AP(Cyclist)/%
0.4～0.5	文中网络	83	82	72
	DETR网络	81	83	71
0.5～0.6	文中网络	82	80	69
	DETR网络	73	82	68
0.6～0.7	文中网络	77	77	64
	DETR网络	62	77	61
0.7～0.8	文中网络	72	74	60
	DETR网络	50	69	53

由表1可知:随着遮挡比的增大,两种方法的检测准确率均出现了下降的趋势;但文中网络在目标物被遮挡的情况下有更好的表现效果,且两种方法的差距随着遮挡比的提升而提升.

不同遮挡物比下各检测网络的平均精度对比,如表2所示.

表2 不同遮挡物比下各检测网络的平均精度对比

Tab.2 Comparison of average accuracy in different detection network with different occlusion proportions

检测方法	$\overline{\text{AP}}$			
	η=0.4～0.5	η=0.5～0.6	η=0.6～0.7	η=0.7～0.8
文中网络	78.99	76.98	72.67	68.67
DETR网络	78.34	74.33	66.67	57.33
YOLOv5s网络	75.10	73.53	70.42	60.51

3 结束语

针对视频中的目标物移动可能产生遮挡、姿势的变化,光照的差异等问题,提出一种基于注意力叠加与时序特征融合的目标检测方法.引入注意力权重叠加的像素级标识提取模块,更好地匹配轨迹.通过已有轨迹信息,采用时序特征融合的方式增强当前帧下的目标检测精度.实验结果证明,文中方法能有效修正目标物被遮挡等情况下的检测效果,是一种具有强鲁棒性的目标检测方法.

参考文献:

[1] 胡珉,周显威,高新闻.公路隧道视频预处理和病害识别算法[J].华侨大学学报(自然科学版),2020,41(5):595-604.DOI:10.11830/ISSN.1000-5013.202002024.

[2] GIRSHICK R,DONAHUE J,DARRELL T,*et al*.Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Columbus:IEEE Press,2014:580-587.DOI:10.1109/CVPR.2014.81.

[3] HE Kaiming,ZHANG Xiangyu,REN Shaoqing,*et al*.Spatial pyramid pooling in deep convolutional networks for visual recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1904-1916.DOI:10.1109/TPAMI.2015.2389824.

[4] REN Shaoqing,HE Kaiming,GIRSHICK R,*et al*.Faster R-CNN: Towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.DOI:10.1109/TPAMI.2016.2577031.

[5] LIU Wei,ANGUELOV D,ERHAN D,*et al*.Ssd: Single shot multibox detector[C]// European Conference on Computer Vision.Amsterdam:Springer,2016:21-37.DOI:10.1007/978-3-319-46448-0_2.

[6] LIN T Y,GOYAL P,GIRSHICK R,*et al*.Focal loss for dense object detection[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,42(2):318-326.

[7] ZHANG Shifeng,WEN Longyin,BIAN Xiao,*et al*.Single-shot refinement neural network for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake City:IEEE Press,2018:4203-4212.DOI:10.1109/CVPR.2018.00442.

[8] REN J,CHEN Xiaohao,LIU Jianbo,*et al*.Accurate single stage detector using recurrent rolling convolution[C]//

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017: 5420-5428. DOI: 10.1109/CVPR.2017.87.

[9] JIAO Licheng, ZHANG Fan, LIU Fang, *et al*. A survey of deep learning-based object detection[J]. IEEE Access, 2019, 7: 128837-128868.

[10] REDMON J, DIVVALA S, GIRSHICK R, *et al*. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Les Vegas: IEEE Press, 2016: 779-788. DOI: 10.1109/CVPR.2016.91.

[11] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017: 7263-7271. DOI: 10.1109/CVPR.2017.690.

[12] TIAN Zhi, SHEN Chunhua, CHEN Hao, *et al*. Fcos: Fully convolutional one-stage object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE Press, 2019: 9627-9636.

[13] CHOI J, CHUN D, KIM H, *et al*. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE Press, 2019: 502-511.

[14] CARION N, MASSA F, SYNNAEVE G, *et al*. End-to-end object detection with transformers[C]// European Conference on Computer Vision. Glasgow: Springer, 2020: 213-229.

[15] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[16] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]// Proceedings of the 31th International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc, 2017: 6000-6010.

[17] ZHU Xizhou, HU Han, LIN S, *et al*. Deformable convnets v2: More deformable, better results[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE Press, 2019: 9308-9316.

[18] XIANG Yu, ALAHI A, SAVARESE S. Learning to track: Online multi-object tracking by decision making[C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE Press, 2015: 4705-4713. DOI: 10.1109/ICCV.2015.534.

[19] 陈柏生，陈锻生. 联合时空特征的车辆跟踪[J]. 华侨大学学报(自然科学版)，2008，29(2)：222-224. DOI：10.11830/ISSN.1000-5013.2008.02.0222.

[20] KANG Kai, OUYANG Wanli, LI Hongsheng, *et al*. Object detection from video tubelets with convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 817-825. DOI: 10.1109/CVPR.2016.95.

[21] CHU Qi, OUYANG Wanli, LI Hongsheng, *et al*. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE Press, 2017: 4836-4845. DOI: 10.1109/ICCV.2017.518.

[22] PANG Bo, LI Yizhuo, ZHANG Yifan, *et al*. Tubetk: Adopting tubes to track multi-object in a one-step training model[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Press, 2020: 6308-6318.

[23] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression[C]// 18th International Conference on Pattern Recognition. Hong Kong: IEEE Press, 2006: 850-855. DOI: 10.1109/ICPR.2006.479.

[24] MNIH V, HEESS N, GRAVES A, *et al*. Recurrent models of visual attention[C]// Advances in Neural Information Processing Systems. New York: Curran Associates, 2014: 2204-2212.

[25] BAHDANAU D, CHOROWSKI J, SERDYUK D, *et al*. End-to-end attention-based large vocabulary speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE Press, 2016: 4945-4949.

（责任编辑：陈志贤　　英文审校：吴逢铁）